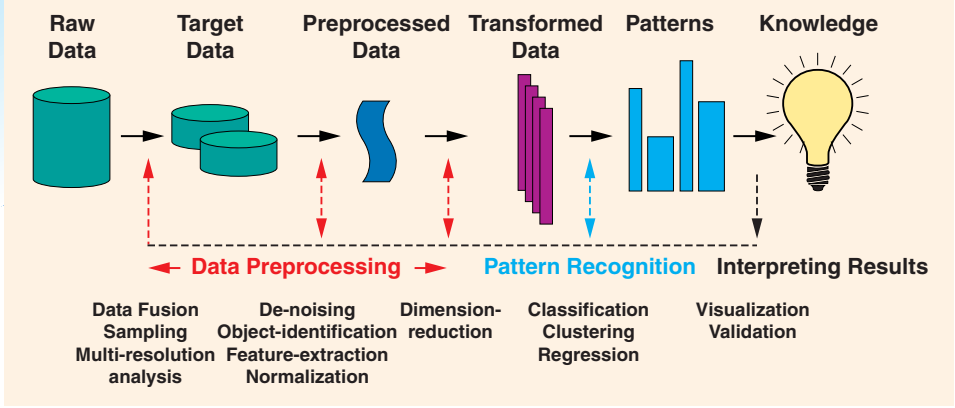
# SAPPHIRE

## Large-scale Data Mining and Pattern Recognition

Center for Applied Scientific Computing

### Mission

The Center for Applied Scientific Computing (CASC) at Lawrence Livermore National Laboratory (LLNL) is developing Sapphire, an object-oriented framework for the interactive exploration of large, complex, multi-dimensional scientific data. By applying and extending ideas from data mining and pattern recognition, we are designing and implementing a new generation of computational tools and techniques that are being used to improve the way in which scientists extract useful information from data.




*Data mining - an iterative and interactive process for finding useful information in massive datasets.*

Advances in technology have enabled scientists to gather data from experiments, simulations, and observations at an ever-increasing pace. Unfortunately, the size and complexity of the data in many scientific domains is such that it is impractical to manually explore, analyze, and understand the data. As a result, useful information is often overlooked, and the potential benefits of increased computational and data-gathering capabilities are only partially realized.

To solve this problem, we are applying and extending ideas from data mining—in particular, pattern recognition—to automatically identify patterns in the data. This makes it possible for scientists to interactively explore the areas of interest in the data.

### Application Domains

The terascale computing environment at Lawrence Livermore National Laboratory (LLNL) has enabled the simulation of increasingly complex phenomena, leading to the generation of vast quantities of data. These simulations play a key role in areas such as nuclear weapons stockpile stewardship, where computer simulations have replaced experiments, and climate modeling, where experiments are impractical or unwise. Visualization techniques are frequently used to help scientists understand the output from these simulations. Often, the size of the data set is such that visualization, by itself, is not sufficient. By coupling visualization with data mining techniques, it becomes possible to allow interactive display of only those areas that are of interest to the scientist, enabling faster exploration of the data. This would not only help in understanding the output from a single simulation, but also help in comparing the output from ensembles of simulations, or assist in comparing simulations with experiments, or controlling the simulations interactively.

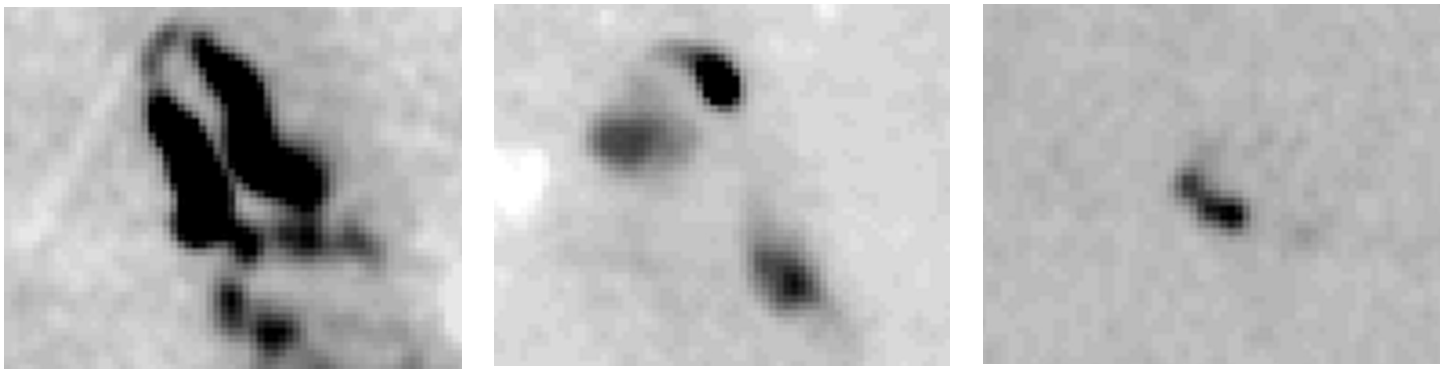
Data mining techniques can also be very useful in domains such as astrophysics, where vast quantities of data are gathered during surveys of the sky. The use of automated techniques can help bring objectivity to the results obtained from these surveys. In addition, data sets that were originally obtained for one purpose can now be reanalyzed to detect previously unknown patterns in the data. This application of knowledge discovery would help address the concerns voiced by astrophysicists that the difficulties in analyzing terabytes of data have resulted in the loss of serendipitous discoveries that were so vital to progress in the area in the past.

### Open Research Problems

Data mining is a process that uncovers patterns, associations, anomalies, and statistically significant structures and events in data. One of the key steps in data mining is pattern recognition, namely, the discovery and characterization of patterns. A pattern is defined as an arrangement or an ordering in which some organization of underlying structure can be said to exist. Patterns in data are identified using measurable features or attributes that have been extracted from the data.

Data mining is an interactive and iterative process involving data pre-processing, search for patterns, knowledge evaluation, and possible refinement of the process based on input from domain experts or feedback from one of the steps. The pre-processing of the data, which is often domain- and application-dependent, is a time-consuming, but critical, first step in the data mining process. The pattern recognition step is usually independent of the domain or application.

Large-scale data mining is a field very much in its infancy, making it a source of several open research problems. In order to extend data mining techniques to large-scale data sets, several barriers must be overcome. The extraction of key features from large, multi-dimensional, complex data is a critical issue that must be addressed



*We are using data mining to identify radio-emitting galaxies with a bent-double morphology.*

first, prior to the application of the pattern recognition algorithms. The features extracted must be relevant to the problem, insensitive to small changes in the data, and invariant to scaling, rotation, and translation. The pattern recognition step poses several challenges as well. For example, is it possible to modify existing algorithms, or design new ones, that are scalable, robust, accurate, and interpretable? Furthermore, can these algorithms be applied effectively and efficiently to complex, multi-dimensional data? Additionally, is it possible to implement these algorithms efficiently on large-scale multiprocessor systems so that a scientist can interactively explore and analyze the data?

Mining scientific data poses additional challenges. For example, data sets from scientific applications are often available as images, posing serious challenges in the extraction of features. Problems in knowledge discovery may be such that the class of interest occurs with low probability, making random sampling inapplicable and traditional clustering techniques ineffective. In addition, high accuracy and precision are required in prediction and description in order to test or refute competing theories. These problems, specific to scientific data sets, preclude the direct application of software and techniques that have been developed for commercial applications.

## Research Directions in Data Mining

Our approach to scaling data mining and pattern recognition algorithms to large, complex, multi-dimensional data addresses each of the steps in the data mining process. Specifically, our research focus includes:

- Image processing techniques for denoising, feature extraction, and object identification.
- Dimension reduction techniques to handle multi-dimensional data.
- Scalable algorithms for classification and clustering.
- Parallel implementations for interactive exploration of data.
- Applied statistics to ensure that the conclusions drawn from the data are statistically sound.

We have designed a flexible object-oriented software infrastructure to implement our algorithms. This enables scientists in a variety of disciplines to experiment with different algorithms, fine-tune an algorithm to a problem, and handle growing datasets. Our research in data pre-processing includes wavelet-based statistical techniques for reducing noise in data, the use of evolutionary algorithms for adaptive image processing, and non-linear, non-orthogonal variants of dimension reduction techniques. We have also explored the creation of ensembles of decision trees by randomizing the decision at each node, and improved the performance of oblique decision trees through the use of evolutionary algorithms.

We are currently working with both observational and simulated data. In collaboration with scientists from the FIRST project (http://sundog.stsci.edu), we have developed algorithms to automatically detect radio-emitting galaxies exhibiting bent-double morphologies. We are now applying and extending our algorithms to the problems of dimension reduction in climate data, the detection of coherent structures in turbulent flows, and the analysis of remotely sensed data.

We expect that our work in these applications will help answer several of the open research questions in the area of data mining and pattern recognition for large, complex, multi-dimensional data. It is our hope that we can successfully address the issue of data overload, and help scientists to explore and understand their data in an effective and efficient manner.

*For additional information about the Sapphire project, see our website at http://www.llnl.gov/casc/sapphire/ or contact Chandrika Kamath, (925) 423-3768, kamath2@llnl.gov.*


This work was performed under the auspices of the U.S. Department of Energy by University of California Lawrence Livermore National Laboratory under contract No. W-7405-Eng-48.